A Review of Machine Learning: Derivation Tables

By: Oameed Noakoasteen
Version 04, Last Updated 2026.08; First Published Online July 2024

View the Presentation on Internet Archive
View the Derivation Tables on Internet Archive
View the GitHub Repository

Oameed Noakoasteen

Table of Contents

Part I: Statistical Learning

1- Supervised Learning: Regression 1

2- Supervised Learning: Classification 4

3- Unsupervised Learning: Dimensionality Reduction 9

4- Ensemble Methods 13

Part II: Deep Learning

5- Artificial Neural Networks 15

Part III: Appendices

Appendix I - A Medley of Concepts 25

Appendix II - Useful Math 27

Oameed Noakoasteen

List of Tables

Table 1. Derivation of optimum weights for Ridge Regression 1
Table 2. Derivation of the cost for Logistic Regression 4
Table 3. Derivation of the cost for Support Vector Classifier 6
Table 4. Derivation of optimal weights for PCA 9
Table 5. Derivation of optimal weights for LDA 10
Table 6. Derivation of the AdaBoost Training Algorithm 13
Table 7. Derivation of the Back-Propagation Algorithm 15
Table 8. Derivation of The He and Glorot Initialization Schemes 18
Table 9. Derivation of the Bias Variance Decomposition for MSE 25
Table 10. Derivation of the mean and variance of predictions in Bayesian inference 26

Oameed Noakoasteen

Notation

■ Data Point

$$\boldsymbol{x}_{n\times 1} \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \tilde{\boldsymbol{x}}_{(n+1)\times 1} \equiv \begin{bmatrix} \boldsymbol{x}_{n\times 1} \\ 1 \end{bmatrix}$$

■ Design Matrix

$$\boldsymbol{X}_{m\times n} \equiv \begin{bmatrix} \boldsymbol{x}_{n\times 1}^{(1)\ T} \\ \vdots \\ \boldsymbol{x}_{n\times 1}^{(m)\ T} \end{bmatrix} \quad \tilde{\boldsymbol{X}}_{m\times (n+1)} \equiv \begin{bmatrix} \tilde{\boldsymbol{x}}_{(n+1)\times 1}^{(1)\ T} \\ \vdots \\ \tilde{\boldsymbol{x}}_{(n+1)\times 1}^{(m)\ T} \end{bmatrix}$$

■ Targets/Labels

$$\boldsymbol{y}_{m\times 1} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \qquad \boldsymbol{Y}_{m\times m} \equiv \begin{bmatrix} y_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & y_m \end{bmatrix}$$

■ Weights & Bias

$$\boldsymbol{w}_{n\times 1} \equiv \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \quad \tilde{\boldsymbol{w}}_{(n+1)\times 1} \equiv \begin{bmatrix} \boldsymbol{w}_{n\times 1} \\ b \end{bmatrix}$$

■ Neuron Weights & Bias

$$\boldsymbol{w}_{O_{l-1}\times 1}^{(l),o} \equiv \begin{bmatrix} w_1^{(l),o} \\ \vdots \\ w_{O_{l-1}}^{(l),o} \end{bmatrix} \qquad \tilde{\boldsymbol{w}}_{(O_{l-1}+1)\times 1}^{(l),o} \equiv \begin{bmatrix} \boldsymbol{w}_{O_{l-1}\times 1}^{(l),o} \\ b^{(l),o} \end{bmatrix}$$

■ Layer Weights & Biases

$$\boldsymbol{W}_{O_l\times O_{l-1}}^{(l)} \equiv \begin{bmatrix} \boldsymbol{w}_{O_{l-1}\times 1}^{(l),1\ T} \\ \vdots \\ \boldsymbol{w}_{O_{l-1}\times 1}^{(l),O_l\ T} \end{bmatrix} \quad \tilde{\boldsymbol{W}}_{O_l\times (O_{l-1}+1)}^{(l)} \equiv \begin{bmatrix} \tilde{\boldsymbol{w}}_{(O_{l-1}+1)\times 1}^{(l),1\ T} \\ \vdots \\ \tilde{\boldsymbol{w}}_{(O_{l-1}+1)\times 1}^{(l),O_l\ T} \end{bmatrix}$$

■ Layer Activations

$$\boldsymbol{a}_{O\times 1}^{(l)} \equiv \begin{bmatrix} a_1^{(l)} \\ \vdots \\ a_O^{(l)} \end{bmatrix} \qquad \tilde{\boldsymbol{a}}_{(O+1)\times 1}^{(l)} \equiv \begin{bmatrix} \boldsymbol{a}_{O\times 1}^{(l)} \\ 1 \end{bmatrix}$$

Oameed Noakoasteen

Oameed Noakoasteen

Part I: Statistical Learning

Oameed Noakoasteen

Oameed Noakoasteen

1- Supervised Learning: Regression

Table 1. Derivation of optimum weights for Ridge Regression

Structural Risk Minimization

The Cost

The loss function is the **squared difference**. The cost is the sum of (a) the average of error over the training examples (which, with this specific loss function, is usually referred to as the **Mean Squared Error**) and (b) a regularization term defined as the norm of the weight vector (usually referred to as the **Ridge Regularizer** or **Weight Decay Regularizer**). (1-a)

$$C(\tilde{\boldsymbol{w}}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{n+1}\tilde{w}_i^2$$

(1-a) → (1-b)

$$= \frac{1}{m}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 + \lambda\|\tilde{\boldsymbol{w}}\|_2^2$$

$$= \frac{1}{m}[(\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}})] + \lambda\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}$$

$$= \frac{1}{m}[\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}^T\boldsymbol{y} + \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}}] + \lambda\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}$$

$$\hat{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}} \quad (2)$$

(1-b), (2) → $$C(\tilde{\boldsymbol{w}}) = \frac{1}{m}\left[\|\boldsymbol{y}\|_2^2 - 2\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}\right] + \lambda\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}} \quad (3)$$

Minimization of the Cost

$$\frac{\partial C(\tilde{\boldsymbol{w}})}{\partial\tilde{\boldsymbol{w}}} = \frac{1}{m}\left[-2\tilde{\boldsymbol{X}}^T\boldsymbol{y} + 2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}\right] + 2\lambda\tilde{\boldsymbol{w}}$$

(3) → (4)

$$= -\frac{2}{m}\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \frac{2}{m}\left[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + m\lambda\boldsymbol{I}\right]\tilde{\boldsymbol{w}}$$

$$= \boldsymbol{0}$$

(4) → $$\tilde{\boldsymbol{w}}^* = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + m\lambda\boldsymbol{I}\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}$$

Oameed Noakoasteen

Posterior Maximization

The Posterior Distribution of $\widetilde{\boldsymbol{w}}$

$$\begin{aligned} p(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}},\boldsymbol{y}) &= \frac{1}{p(\widetilde{\boldsymbol{X}},\boldsymbol{y})}p(\widetilde{\boldsymbol{X}},\boldsymbol{y}|\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{w}}) \\ &= \frac{1}{p(\boldsymbol{y}|\widetilde{\boldsymbol{X}})p(\widetilde{\boldsymbol{X}})}p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{X}})p(\widetilde{\boldsymbol{w}}) \\ &= \frac{1}{p(\boldsymbol{y}|\widetilde{\boldsymbol{X}})}p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{w}}) \end{aligned} \tag{5-a}$$

The expression $\frac{1}{p(\boldsymbol{y}|\widetilde{\boldsymbol{X}})}$ in (5-a) is not a function of $\widetilde{\boldsymbol{w}}$ and is, therefore, only a constant scaling factor. For this reason, only the expression $p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{w}})$ is considered in the following derivations. (5-b)

(5-a), (5-b) → $$p(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}},\boldsymbol{y}) \propto p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{w}}) \tag{5-c}$$

The Prior Distribution of $\widetilde{\boldsymbol{w}}$

The random vector $\widetilde{\boldsymbol{w}}$ is drawn from a zero-mean multivariate Gaussian distribution $\widetilde{\boldsymbol{w}}\sim\mathcal{N}\left(\boldsymbol{\mu}=\mathbf{0}_{(n+1)\times 1},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{prior}_{(n+1)\times(n+1)}\right)$. (6)

(6) → $$p(\widetilde{\boldsymbol{w}}) = \frac{1}{(2\pi)^{\frac{n+1}{2}}|\boldsymbol{\Sigma}^{prior}|^{\frac{1}{2}}}e^{-\frac{1}{2}\widetilde{\boldsymbol{w}}^T\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\widetilde{\boldsymbol{w}}} \tag{7}$$

The Likelihood of $\widetilde{\boldsymbol{w}}$

The true targets are related to their predicted values by $\boldsymbol{y}=\widehat{\boldsymbol{y}}+\boldsymbol{\varepsilon}$ in which the random vector $\boldsymbol{\varepsilon}$ is drawn from a zero-mean multivariate Gaussian distribution $\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{\mu}=\mathbf{0}_{m\times 1},\boldsymbol{\Sigma}=\sigma^2\boldsymbol{I}_{m\times m})$. (8)

(2), (8) →
$$\begin{aligned} p(\boldsymbol{\varepsilon}) &= \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}}e^{-\frac{1}{2}\boldsymbol{\varepsilon}^T(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{\varepsilon}} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}}e^{-\frac{1}{2}(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}})} \\ &= p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}}) \end{aligned} \tag{9}$$

(5-c), (7), (9) →
$$\begin{aligned} p(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}},\boldsymbol{y}) &\propto p(\boldsymbol{y}|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{w}})p(\widetilde{\boldsymbol{w}}) \\ &\propto e^{-\frac{1}{2}\left[(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}})+\widetilde{\boldsymbol{w}}^T\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\widetilde{\boldsymbol{w}}\right]} \\ &\propto e^{-\frac{1}{2\sigma^2}\left[\widetilde{\boldsymbol{w}}^T\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}+\sigma^2\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\right)\widetilde{\boldsymbol{w}}-2\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{X}}^T\boldsymbol{y}\right]} \end{aligned} \tag{10}$$

Oameed Noakoasteen

In equation (10), defining $\boldsymbol{A} \equiv \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\right)^{-1}$ and also $\boldsymbol{b} \equiv \widetilde{\boldsymbol{X}}^T\boldsymbol{y}$ and then adding and subtracting $\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}$ (which is not a function of $\widetilde{\boldsymbol{w}}$) to complete the square (plus, noting that $\boldsymbol{A}^T = \boldsymbol{A}$), results in $p\left(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}}, \boldsymbol{y}\right) \propto e^{-\frac{1}{2}(\widetilde{\boldsymbol{w}}-\boldsymbol{A}\boldsymbol{b})^T(\sigma^2\boldsymbol{A})^{-1}(\widetilde{\boldsymbol{w}}-\boldsymbol{A}\boldsymbol{b})}$. (11)

(10), (11) → $$p\left(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}}, \boldsymbol{y}\right) = \mathcal{N}(\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{b}, \boldsymbol{\Sigma} = \sigma^2\boldsymbol{A}) \quad (12)$$

The Cost

The cost is the **negative-log-of-the-posterior**. (13-a)

$$\begin{aligned} C(\widetilde{\boldsymbol{w}}) &= -\log\left(p\left(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}}, \boldsymbol{y}\right)\right) \\ &= \frac{1}{2}(\widetilde{\boldsymbol{w}} - \boldsymbol{A}\boldsymbol{b})^T(\sigma^2\boldsymbol{A})^{-1}(\widetilde{\boldsymbol{w}} - \boldsymbol{A}\boldsymbol{b}) \\ &= \frac{1}{2\sigma^2}[\widetilde{\boldsymbol{w}}^T\boldsymbol{A}^{-1}\widetilde{\boldsymbol{w}} - 2\widetilde{\boldsymbol{w}}^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}] \end{aligned}$$ (12), (13-a) → (13-b)

Minimization of the Cost

$$\begin{aligned} \frac{\partial C(\widetilde{\boldsymbol{w}})}{\partial \widetilde{\boldsymbol{w}}} &= \frac{1}{2\sigma^2}[2\boldsymbol{A}^{-1}\widetilde{\boldsymbol{w}} - 2\boldsymbol{b}] \\ &= \boldsymbol{0} \end{aligned}$$ (13-b) → (14)

(11), (14) → $$\widetilde{\boldsymbol{w}}^* = \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\right)^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$$

The Dual Formulation of $\widetilde{\boldsymbol{w}}^*$

$$\widetilde{\boldsymbol{w}}^* = \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + k\boldsymbol{I}\right)^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y} \quad (15)$$

Desired Transformation According to the Representer Theorem

$$\widetilde{\boldsymbol{w}}^* = \widetilde{\boldsymbol{X}}^T\boldsymbol{\alpha} \quad (16)$$

(15), (16) → $$\boldsymbol{y} = \left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + k\boldsymbol{I}\right)\boldsymbol{\alpha} \quad (17)$$

(17) → $$\boldsymbol{\alpha} = \left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + k\boldsymbol{I}\right)^{-1}\boldsymbol{y}$$

Oameed Noakoasteen

2- Supervised Learning: Classification

Table 2. Derivation of the cost for Logistic Regression

Posterior Maximization

The Posterior Distribution of $\widetilde{\boldsymbol{w}}$

The Prior Distribution of $\widetilde{\boldsymbol{w}}$

The random vector $\widetilde{\boldsymbol{w}}$ is drawn from a zero-mean multivariate Gaussian distribution $\widetilde{\boldsymbol{w}} \sim \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{0}_{(n+1)\times 1}, \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{(n+1)\times(n+1)}\right)$. (1-a)

(1-a) → $$p(\widetilde{\boldsymbol{w}}) = \frac{1}{(2\pi\sigma^2)^{\frac{n+1}{2}}} e^{-\frac{1}{2\sigma^2}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{w}}} \quad (1\text{-b})$$

The Likelihood of $\widetilde{\boldsymbol{w}}$

The Logistic (Sigmoid) Function

$$\sigma(z) \equiv \frac{1}{1+e^{-z}} \quad (2\text{-a})$$

(2-a) → $$p(y=1|\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{w}}) \equiv \sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}) \quad (2\text{-b})$$

(2-b) → $$p(y|\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{w}}) = [p(y=1|\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{w}})]^y [1 - p(y=1|\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{w}})]^{1-y} \quad (2\text{-c})$$

(2-c) → $$p(\boldsymbol{y}|\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{w}}) = \prod_{i=1}^{m} p\left(y_i|\widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{w}}\right) \quad (2\text{-d})$$

(1-b), (2) → $$\begin{aligned} p(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}}, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{w}}) p(\widetilde{\boldsymbol{w}}) \\ &\propto \left(\prod_{i=1}^{m} p\left(y_i|\widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{w}}\right)\right) e^{-\frac{1}{2\sigma^2}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{w}}} \end{aligned} \quad (3)$$

The Cost

The cost is the **negative-log-of-the-posterior**. (4-a)

(3), (4-a) → $$\begin{aligned} C(\widetilde{\boldsymbol{w}}) &= -\frac{1}{m}\log\left(p(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}}, \boldsymbol{y})\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} -\log\left(p\left(y_i|\widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{w}}\right)\right) + \frac{1}{2m\sigma^2}\|\widetilde{\boldsymbol{w}}\|_2^2 \end{aligned} \quad (4\text{-b})$$

Oameed Noakoasteen

Minimization of the Cost

The Gradients

$$\frac{\partial C(\tilde{\boldsymbol{w}})}{\partial \tilde{w}_j} = -\frac{1}{m}\sum_{i=1}^{m} y_i \frac{\partial}{\partial \tilde{w}_j}\left(\log\left(\sigma\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}^{(i)}\right)\right)\right) + -\frac{1}{m}\sum_{i=1}^{m}(1-y_i)\frac{\partial}{\partial \tilde{w}_j}\left(\log\left(1-\sigma\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}^{(i)}\right)\right)\right) + \frac{1}{m\sigma^2}\tilde{w}_j$$

(4-b) →

$$= -\frac{1}{m}\sum_{i=1}^{m} y_i\tilde{x}_j^{(i)}\left(1-\sigma\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}^{(i)}\right)\right) + -\frac{1}{m}\sum_{i=1}^{m} -(1-y_i)\tilde{x}_j^{(i)}\sigma\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}^{(i)}\right) + \frac{1}{m\sigma^2}\tilde{w}_j$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\sigma\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}^{(i)}\right)-y_i\right)\tilde{x}_j^{(i)} + \frac{1}{m\sigma^2}\tilde{w}_j$$

Oameed Noakoasteen

Table 3. Derivation of the cost for Support Vector Classifier

Structural Risk Minimization

The Concepts Involved in Construction of the Cost

The Estimating Hyper-Plane

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0$$

The plane f is defined by a point $\boldsymbol{x}_0$ that lies on it and a vector $\boldsymbol{w}$ that is normal to it. A given point $\boldsymbol{x}$ lies on plane f if the plane's normal vector $\boldsymbol{w}$ and the vector defined by points $\boldsymbol{x}_0$ and $\boldsymbol{x}$, i.e., $\Delta\boldsymbol{x}$ are perpendicular. Therefore, the equation for the plane is given by $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \Delta\boldsymbol{x}$ and the point $\boldsymbol{x}$ lies on it if $f(\boldsymbol{x}) = 0$.

The Margins

The $+1$ and the -1 margins are defined as the locus of $\boldsymbol{x}$ for which $|\boldsymbol{w}^T\boldsymbol{x} + b| = 1$.

Since the distance between any point $\boldsymbol{x}$ and the plane f is given by $d = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$, the distance between the margins and the hyperplane is given by $d = \frac{1}{\|\boldsymbol{w}\|}$.

The equation of plane f can also be expressed as $f(\boldsymbol{x}) = \|\boldsymbol{w}\|\|\Delta\boldsymbol{x}\| \cos\theta$ in which θ is the angle between $\boldsymbol{w}$ and $\Delta\boldsymbol{x}$. The expression $\|\Delta\boldsymbol{x}\| \cos\theta$ is the (signed) distance between point $\boldsymbol{x}$ and plane f. Therefore, $d = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$.

The Hinge Loss Function

$$L_H\big(y, f(\boldsymbol{x})\big) \equiv \max\big(0, 1 - yf(\boldsymbol{x})\big)$$

For any example $(\boldsymbol{x}, y)$, If data-point $\boldsymbol{x}$ is correctly classified and is away from or on the margins, i.e., $y(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$, then $L_H = 0$. If data-point $\boldsymbol{x}$ is correctly classified but is inside the margins or it is incorrectly classified, i.e., $y(\boldsymbol{w}^T\boldsymbol{x} + b) < 1$, then $L_H > 0$. (1)

The Slack Variable

For any example $(\boldsymbol{x}, y)$, $\xi = 0$ if $L_H = 0$ and $\xi \geq L_H$ if $L_H > 0$. (2)

Oameed Noakoasteen

The Cost

The cost is the sum of (a) the average of slack variables over the training examples and (b) a regularization term defined as the norm of the weight vector (since minimization of the $\|\boldsymbol{w}\|$ results in maximization of the distance between the hyperplane and the margins which in turn reduces/moderates the model's capacity/**VC Dimension**). (3)

Minimization of the Cost

$$\min_{\boldsymbol{w},b}\left(C'\frac{1}{m}\sum_{i=1}^{m}\xi_i+\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\right)$$

(3) → subject to (4)

$$\xi_i \geq 1-y_i\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)$$
$$\xi_i \geq 0$$

The Method of Lagrange Multipliers

The Primal Form of the Lagrangian

$$L_p = C\sum_{i=1}^{m}\xi_i+\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{m}\alpha_i\left(\xi_i-1+y_i\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right) - \sum_{i=1}^{m}\mu_i\xi_i \quad (5\text{-a})$$

(4) → subject to (Karush-Kuhn-Tucker Conditions)

$$\alpha_i \geq 0 \quad (5\text{-b})$$
$$\mu_i \geq 0 \quad (5\text{-c})$$
$$\alpha_i\left(\xi_i-1+y_i\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right)=0 \quad (5\text{-d})$$
$$\mu_i\xi_i=0 \quad (5\text{-e})$$

The Dual Form of the Lagrangian

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w}-\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}^{(i)} = \boldsymbol{w}-\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\alpha} = \boldsymbol{0} \quad (6)$$

Oameed Noakoasteen

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \quad (7)$$

$$\frac{\partial L_p}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad (8)$$

$$\boldsymbol{w} = \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{\alpha} \quad (9\text{-a})$$

(6), (7), (8) → $\sum_{i=1}^{m} \alpha_i y_i = 0$ (9-b)

$\alpha_i + \mu_i = C$ (9-c)

(5-a), (9) → $L_d = -\frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{Y} \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{\alpha} + \boldsymbol{\alpha} \cdot \mathbf{1}$

(1), (2), (5-d), (5-e), (9-c) → For any example $(\boldsymbol{x}, y)$, $\alpha = 0$ if it is correctly classified and is away from the margins; $\alpha = C$ if it is inside the margins or is incorrectly classified. Therefore, $0 < \alpha < C$ if it lies on any of the margins.

Oameed Noakoasteen

3- Unsupervised Learning: Dimensionality Reduction

Table 4. Derivation of optimal weights for PCA

The transformation $\boldsymbol{W}_{n\times k}^{T}$ maps each datapoint $\boldsymbol{x}_{n\times 1}$ to a lower-dimensional representation $\boldsymbol{z}_{k\times 1}$, i.e., $\boldsymbol{z}_{k\times 1}=\boldsymbol{W}_{n\times k}^{T}\boldsymbol{x}_{n\times 1}$. In matrix form, $\boldsymbol{Z}_{m\times k}=\boldsymbol{X}_{m\times n}\boldsymbol{W}_{n\times k}$. Furthermore, the original datapoint is reconstructed by $\widehat{\boldsymbol{x}}_{n\times 1}=\boldsymbol{W}_{n\times k}\boldsymbol{z}_{k\times 1}$. In matrix form, $\widehat{\boldsymbol{X}}_{m\times n}=\boldsymbol{Z}_{m\times k}\boldsymbol{W}_{n\times k}^{T}$. (1-a)

There are two equivalent guiding principles for obtaining $\boldsymbol{W}_{n\times k}^{*}$:

- Variance of the mappings is maximized, i.e., $\boldsymbol{W}^{*}=\underset{\boldsymbol{W}}{\operatorname{argmax}}\,Tr(K_{\boldsymbol{zz}})$
- The Reconstruction Error is minimized, i.e., $\boldsymbol{W}^{*}=\underset{\boldsymbol{W}}{\operatorname{argmin}}\,RE\left(\boldsymbol{X},\widehat{\boldsymbol{X}}\right)$

(1-b)

Furthermore, it is required that $\boldsymbol{W}\boldsymbol{W}^{T}=\boldsymbol{I}_{n\times n}$.

Useful Math

The Trace operation is invariant to cyclic permutation even if changing the order of multiplications results in matrices of different dimensions, i.e., $Tr\left(\boldsymbol{A}_{p\times q}\boldsymbol{B}_{q\times p}\right)=Tr\left(\boldsymbol{B}_{q\times p}\boldsymbol{A}_{p\times q}\right)$. I-1

Variance Maximization

Variance of Mappings

(1-a) →

$$\begin{aligned} K_{\boldsymbol{zz}} &\equiv E[(\boldsymbol{z}-E[\boldsymbol{z}])(\boldsymbol{z}-E[\boldsymbol{z}])^{T}] \\ &= E[\boldsymbol{z}\boldsymbol{z}^{T}]-E[\boldsymbol{z}]E^{T}[\boldsymbol{z}] \\ &= \boldsymbol{W}^{T}E[\boldsymbol{x}\boldsymbol{x}^{T}]\boldsymbol{W}-\boldsymbol{W}^{T}E[\boldsymbol{x}]E^{T}[\boldsymbol{x}]\boldsymbol{W} \\ &= \boldsymbol{W}^{T}(E[\boldsymbol{x}\boldsymbol{x}^{T}]-E[\boldsymbol{x}]E^{T}[\boldsymbol{x}])\boldsymbol{W} \\ &= \boldsymbol{W}^{T}K_{\boldsymbol{xx}}\boldsymbol{W} \\ &= \frac{1}{m}\boldsymbol{W}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{W} \end{aligned} \tag{2}$$

From (2) it follows that, $\boldsymbol{W}^{*}$ must maximize $Tr(\boldsymbol{W}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{W})$. This can be achieved by constructing $\boldsymbol{W}$ using the k eigenvectors of $\boldsymbol{X}^{T}\boldsymbol{X}$ corresponding to its k largest eigenvalues.

Reconstruction Error Minimization

The Reconstruction Error

(1), (I-1) →

$$\begin{aligned} RE\left(\boldsymbol{X},\widehat{\boldsymbol{X}}\right) &\equiv \left\|\boldsymbol{X}-\widehat{\boldsymbol{X}}\right\|_{F}^{2} \\ &= Tr\left(\left(\boldsymbol{X}-\widehat{\boldsymbol{X}}\right)\left(\boldsymbol{X}-\widehat{\boldsymbol{X}}\right)^{T}\right) \\ &= Tr\left(\boldsymbol{X}\boldsymbol{X}^{T}-\boldsymbol{X}\widehat{\boldsymbol{X}}^{T}-\widehat{\boldsymbol{X}}\boldsymbol{X}^{T}+\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^{T}\right) \\ &= Tr(\boldsymbol{X}\boldsymbol{X}^{T}-\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{X}^{T}-\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{X}^{T}+\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{X}^{T}) \end{aligned} \tag{3}$$

Oameed Noakoasteen

$$
\begin{aligned}
&= Tr(\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}^T) \\
&= Tr(\boldsymbol{X}\boldsymbol{X}^T) - Tr(\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}^T) \\
&= Tr(\boldsymbol{X}\boldsymbol{X}^T) - Tr(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W})
\end{aligned}
$$

The expression in (3) is minimized by maximization of $Tr(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W})$. This can be achieved by constructing $\boldsymbol{W}$ using the k eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ corresponding to its k largest eigenvalues.

Table 5. Derivation of optimal weights for LDA

The dataset $\boldsymbol{X}_{m\times n}$ comprises two separate classes, i.e., $\boldsymbol{X}_{m_0\times n}$ and $\boldsymbol{X}_{m_1\times n}$ such that $m_0 + m_1 = m$. The transformation $\boldsymbol{w}_{n\times 1}^T$ maps each datapoint to a scalar, i.e., $z = \boldsymbol{w}_{n\times 1}^T\boldsymbol{x}_{n\times 1}$. In vector form, for class $c \in \{0,1\}$, $\boldsymbol{z}_{m_c\times 1} = \boldsymbol{X}_{m_c\times n}\boldsymbol{w}_{n\times 1}$. (1-a)

The guiding principle for obtaining $\boldsymbol{w}_{n\times 1}^*$ is that the mappings must simultaneously satisfy two conditions: (a) the distance between the means of the mapped classes, denoted by d_B, must be maximum and (b) the sum of the two within-class variances, denoted by d_W, must be minimum. Therefore, $\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{d_B(\boldsymbol{w})}{d_W(\boldsymbol{w})}$. (1-b)

Useful Math

The Trace operation is invariant to cyclic permutation even if changing the order of multiplications results in matrices of different dimensions, i.e., $Tr(\boldsymbol{A}_{p\times q}\boldsymbol{B}_{q\times p}) = Tr(\boldsymbol{B}_{q\times p}\boldsymbol{A}_{p\times q})$. I-1

The Distance Between the Means of Projected Classes

The Mean of Datapoints

$$\boldsymbol{\mu}_c \equiv \frac{1}{m_c}\boldsymbol{X}_{m_c\times n}^T\mathbf{1}_{m_c\times 1} \tag{2-a}$$

The Mean of Projections

$$
\begin{aligned}
\bar{z}_c &\equiv \frac{1}{m_c}\boldsymbol{z}_{m_c\times 1}^T\mathbf{1}_{m_c\times 1} \\
\text{(2-a)} \rightarrow \quad &= \frac{1}{m_c}\boldsymbol{w}_{n\times 1}^T\boldsymbol{X}_{m_c\times 1}^T\mathbf{1}_{m_c\times 1} \\
&= \boldsymbol{w}_{n\times 1}^T\boldsymbol{\mu}_c
\end{aligned}
\tag{2-b}
$$

Omeed Noakoasteen

The Distance Between the Means

$$\begin{aligned} d_B &\equiv \|\bar{z_0} - \bar{z_1}\|_2^2 \\ &= \|\boldsymbol{w}^T\boldsymbol{\mu}_0 - \boldsymbol{w}^T\boldsymbol{\mu}_1\|_2^2 \\ &= (\boldsymbol{w}^T\boldsymbol{\mu}_0 - \boldsymbol{w}^T\boldsymbol{\mu}_1)^T(\boldsymbol{w}^T\boldsymbol{\mu}_0 - \boldsymbol{w}^T\boldsymbol{\mu}_1) \\ &= (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\ &= Tr\big((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\big) \\ &= Tr(\boldsymbol{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w}) \\ &= \boldsymbol{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w} \end{aligned}$$

(I-1), (2-b) → (2-c)

(2-c) → $d_B = \boldsymbol{w}^T S_B \boldsymbol{w}$ in which $S_B \equiv (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T$. (3)

The Sum of the Two Within-Class Variances

Within-Class Variance

$$\begin{aligned} Var(z) &\equiv E[(z - E[z])(z - E[z])] \\ &= E[z^2] - E^2[z] \\ &= E[zz^T] - E[z]E^T[z] \\ &= E[\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w}] - E[\boldsymbol{w}^T\boldsymbol{x}]E^T[\boldsymbol{w}^T\boldsymbol{x}] \\ &= \boldsymbol{w}^T E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w} - \boldsymbol{w}^T E[\boldsymbol{x}]E^T[\boldsymbol{x}]\boldsymbol{w} \\ &= \boldsymbol{w}^T(E[\boldsymbol{x}\boldsymbol{x}^T] - E[\boldsymbol{x}]E^T[\boldsymbol{x}])\boldsymbol{w} \\ &= \boldsymbol{w}^T K_{\boldsymbol{xx}}\boldsymbol{w} \end{aligned}$$ (4-a)

From (4-a) it follows that, the within-class variance is given by $Var(z) = \boldsymbol{w}^T\boldsymbol{\Sigma}_c\boldsymbol{w}$ in which $\boldsymbol{\Sigma}_c$ denotes $K_{\boldsymbol{xx}}$ for datapoints of class c. (4-b)

The sum

$$\begin{aligned} d_W &\equiv \boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w} \\ &= \boldsymbol{w}^T(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w} \end{aligned}$$

(4-b) → (4-c)

(4-c) → $d_W = \boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}$ in which $\boldsymbol{S}_W \equiv \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$. (5)

Maximization of the $\frac{d_B}{d_W}$

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}$$

(1-b), (3), (5) → subject to

$$\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w} = 1$$

The Method of Lagrange Multipliers

The Primal Form of the Lagrangian

$$L_p \equiv \boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w} - \alpha(\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w} - 1)$$

Oameed Noakoasteen

$$\frac{\partial L_d}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_B\boldsymbol{w} - 2\alpha\boldsymbol{S}_W\boldsymbol{w} = \boldsymbol{0} \quad (6)$$

(6) $\rightarrow$ $$\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{w} = \alpha\boldsymbol{w} \quad (7)$$

From (7) it follows that, $\boldsymbol{w}^*$ must be an eigenvector of $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$.

Oameed Noakoasteen

4- Ensemble Methods

Table 6. Derivation of the AdaBoost Training Algorithm

- The ensemble is defined as $F(\boldsymbol{x}) \equiv \sum_j \alpha_j f_j(\boldsymbol{x})$ in which $f_j(\boldsymbol{x}) \in \{-1, +1\}$ and $F(\boldsymbol{x}) \in \{-1, +1\}$. The loss function is defined as $L(F(\boldsymbol{x}), y) \equiv e^{-yF(\boldsymbol{x})}$ and, therefore, the cost is given by $C(F) = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i F(\boldsymbol{x}^{(i)})}$.
- The next weak learner to be added to the ensemble must be chosen such that the cost is minimized, i.e., $f^* = \underset{f}{\operatorname{argmin}}\, C(F + \alpha f)$. Once the weak learner is chosen, the step size must be determined such that the cost is minimized, i.e., $\alpha^* = \underset{\alpha}{\operatorname{argmin}}\, C(F + \alpha f)$.

(1)

Useful Math

Tylor Approximation

$$C(F + \alpha f) \approx C(F) + \alpha \sum_{i=1}^{m} \frac{\partial C}{\partial F(\boldsymbol{x}^{(i)})} f(\boldsymbol{x}^{(i)}) \qquad \text{I-1}$$

A Probability Distribution Over the Training Set

$$\boldsymbol{w}_{m\times 1} \equiv \text{softmax}(\boldsymbol{y}_{m\times 1} \odot \boldsymbol{F}_{m\times 1})$$

The "softmax" function accepts vector $\boldsymbol{z}$ as input and produces vector $\boldsymbol{z'}$ as output such that the k^{th} element of the output is given by $z'_k = \frac{e^{-z_k}}{\sum_i e^{-z_i}}$.

(2)

As a result of this definition, $\sum_{i=1}^{m} w_i = 1$.

The Optimal Weak Learner

$$\begin{aligned} f^* &= \underset{f}{\operatorname{argmin}}\, C(F + \alpha f) \\ &= \underset{f}{\operatorname{argmin}} \sum_{i=1}^{m} \frac{\partial C}{\partial F(\boldsymbol{x}^{(i)})} f(\boldsymbol{x}^{(i)}) \end{aligned} \qquad (3)$$

(1), I-1 →

In which $\sum_{i=1}^{m} \frac{\partial C}{\partial F(\boldsymbol{x}^{(i)})} f(\boldsymbol{x}^{(i)}) < 0$.

Oameed Noakoasteen

$$
\begin{aligned}
& \sum_{i=1}^{m} \frac{\partial C}{\partial F(\boldsymbol{x}^{(i)})} f(\boldsymbol{x}^{(i)}) \\
= & -\frac{1}{m} \sum_{i=1}^{m} y_i e^{-y_i F(\boldsymbol{x}^{(i)})} f(\boldsymbol{x}^{(i)}) \\
= & -\frac{\sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \sum_{i=1}^{m} w_i y_i f(\boldsymbol{x}^{(i)}) \\
(1), (2) \quad \rightarrow \quad = & \frac{\sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \left[\sum_{i: f(\boldsymbol{x}^{(i)}) \neq y_i} w_i - \sum_{i: f(\boldsymbol{x}^{(i)}) = y_i} w_i \right] \\
= & \frac{\sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \left[2 \sum_{i: f(\boldsymbol{x}^{(i)}) \neq y_i} w_i - 1 \right] \\
= & -\frac{\sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} + \frac{2 \sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \sum_{i=1}^{m} w_i \mathbb{I}_{f(\boldsymbol{x}^{(i)}) \neq y_i}(\boldsymbol{x}^{(i)})
\end{aligned}
\tag{4}
$$

$$
(3), (4) \quad \rightarrow \quad f^* = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{m} w_i \mathbb{I}_{f(\boldsymbol{x}^{(i)}) \neq y_i}(\boldsymbol{x}^{(i)})
$$

The Optimal Step Size

$$
\begin{aligned}
& \frac{\partial C(F + \alpha f)}{\partial \alpha} \\
= & \frac{\partial}{\partial \alpha} \left(\frac{1}{m} \sum_{i=1}^{m} e^{-y_i \left(F(\boldsymbol{x}^{(i)}) + \alpha f(\boldsymbol{x}^{(i)}) \right)} \right) \\
= & -\frac{1}{m} \sum_{i=1}^{m} y_i f(\boldsymbol{x}^{(i)}) e^{-y_i \left(F(\boldsymbol{x}^{(i)}) + \alpha f(\boldsymbol{x}^{(i)}) \right)} \\
= & -\frac{1}{m} \sum_{i: f(\boldsymbol{x}^{(i)}) = y_i} e^{-\alpha} e^{-y_i F(\boldsymbol{x}^{(i)})} + \frac{1}{m} \sum_{i: f(\boldsymbol{x}^{(i)}) \neq y_i} e^{\alpha} e^{-y_i F(\boldsymbol{x}^{(i)})} \\
= & -\frac{e^{-\alpha} \sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \left(1 - \sum_{i: f(\boldsymbol{x}^{(i)}) \neq y_i} w_i \right) + \frac{e^{\alpha} \sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})}}{m} \left(\sum_{i: f(\boldsymbol{x}^{(i)}) \neq y_i} w_i \right) \\
= & \sum_{l=1}^{m} e^{-y_l F(\boldsymbol{x}^{(l)})} \left[-\frac{e^{-\alpha}}{m} + \left(\frac{e^{-\alpha}}{m} + \frac{e^{\alpha}}{m} \right) \sum_{i=1}^{m} w_i \mathbb{I}_{f(\boldsymbol{x}^{(i)}) \neq y_i}(\boldsymbol{x}^{(i)}) \right] \\
= & \ 0
\end{aligned}
\tag{5}
$$

$$
\epsilon \equiv \sum_{i=1}^{m} w_i \mathbb{I}_{f(\boldsymbol{x}^{(i)}) \neq y_i}(\boldsymbol{x}^{(i)}) \tag{6}
$$

$$
(5), (6) \quad \rightarrow \quad \alpha = \frac{1}{2} \ln \left(\frac{1 - \epsilon}{\epsilon} \right)
$$

Oameed Noakoasteen

Part II: Deep Learning

Oameed Noakoasteen

Oameed Noakoasteen

5- Artificial Neural Networks

Table 7. Derivation of the Back-Propagation Algorithm

Layer Three

$$\frac{\partial C}{\partial \widetilde{\boldsymbol{W}}^{(3)}} \equiv \begin{bmatrix} \frac{\partial C}{\partial W_{11}^{(3)}} & \frac{\partial C}{\partial W_{12}^{(3)}} & \frac{\partial C}{\partial W_{13}^{(3)}} \\ \frac{\partial C}{\partial W_{21}^{(3)}} & \frac{\partial C}{\partial W_{22}^{(3)}} & \frac{\partial C}{\partial W_{23}^{(3)}} \\ \frac{\partial C}{\partial W_{31}^{(3)}} & \frac{\partial C}{\partial W_{32}^{(3)}} & \frac{\partial C}{\partial W_{33}^{(3)}} \\ \frac{\partial C}{\partial W_{41}^{(3)}} & \frac{\partial C}{\partial W_{42}^{(3)}} & \frac{\partial C}{\partial W_{43}^{(3)}} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{\partial C}{\partial a_1^{(3)}}\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)a_1^{(2)} & \left(\frac{\partial C}{\partial a_1^{(3)}}\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)a_2^{(2)} & \left(\frac{\partial C}{\partial a_1^{(3)}}\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)1 \\ \left(\frac{\partial C}{\partial a_2^{(3)}}\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)a_1^{(2)} & \left(\frac{\partial C}{\partial a_2^{(3)}}\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)a_2^{(2)} & \left(\frac{\partial C}{\partial a_2^{(3)}}\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)1 \\ \left(\frac{\partial C}{\partial a_3^{(3)}}\frac{\partial a_3^{(3)}}{\partial z_3^{(3)}}\right)a_1^{(2)} & \left(\frac{\partial C}{\partial a_3^{(3)}}\frac{\partial a_3^{(3)}}{\partial z_3^{(3)}}\right)a_2^{(2)} & \left(\frac{\partial C}{\partial a_3^{(3)}}\frac{\partial a_3^{(3)}}{\partial z_3^{(3)}}\right)1 \\ \left(\frac{\partial C}{\partial a_4^{(3)}}\frac{\partial a_4^{(3)}}{\partial z_4^{(3)}}\right)a_1^{(2)} & \left(\frac{\partial C}{\partial a_4^{(3)}}\frac{\partial a_4^{(3)}}{\partial z_4^{(3)}}\right)a_2^{(2)} & \left(\frac{\partial C}{\partial a_4^{(3)}}\frac{\partial a_4^{(3)}}{\partial z_4^{(3)}}\right)1 \end{bmatrix}$$

$$= \left(\boldsymbol{f}'^{(3)} \odot \boldsymbol{\delta}^{(3)}\right)\widetilde{\boldsymbol{a}}^{(2)^T}$$

$$\boldsymbol{f}'^{(3)} \equiv \begin{bmatrix} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \\ \frac{\partial a_2^{(3)}}{\partial z_2^{(3)}} \\ \frac{\partial a_3^{(3)}}{\partial z_3^{(3)}} \\ \frac{\partial a_4^{(3)}}{\partial z_4^{(3)}} \end{bmatrix}$$

Oameed Noakoasteen

$$\boldsymbol{\delta}^{(3)} \equiv \begin{bmatrix} \dfrac{\partial C}{\partial a_1^{(3)}} \\ \dfrac{\partial C}{\partial a_2^{(3)}} \\ \dfrac{\partial C}{\partial a_3^{(3)}} \\ \dfrac{\partial C}{\partial a_4^{(3)}} \end{bmatrix}$$

Layer Two

$$\frac{\partial C}{\partial \widetilde{\boldsymbol{W}}^{(2)}} \equiv \begin{bmatrix} \dfrac{\partial C}{\partial W_{11}^{(2)}} & \dfrac{\partial C}{\partial W_{12}^{(2)}} & \dfrac{\partial C}{\partial W_{13}^{(2)}} & \dfrac{\partial C}{\partial W_{14}^{(2)}} \\ \dfrac{\partial C}{\partial W_{21}^{(2)}} & \dfrac{\partial C}{\partial W_{22}^{(2)}} & \dfrac{\partial C}{\partial W_{23}^{(2)}} & \dfrac{\partial C}{\partial W_{24}^{(2)}} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\dfrac{\partial C}{\partial a_1^{(2)}}\dfrac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)a_1^{(1)} & \left(\dfrac{\partial C}{\partial a_1^{(2)}}\dfrac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)a_2^{(1)} & \left(\dfrac{\partial C}{\partial a_1^{(2)}}\dfrac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)a_3^{(1)} & \left(\dfrac{\partial C}{\partial a_1^{(2)}}\dfrac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)1 \\ \left(\dfrac{\partial C}{\partial a_2^{(2)}}\dfrac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)a_1^{(1)} & \left(\dfrac{\partial C}{\partial a_2^{(2)}}\dfrac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)a_2^{(1)} & \left(\dfrac{\partial C}{\partial a_2^{(2)}}\dfrac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)a_3^{(1)} & \left(\dfrac{\partial C}{\partial a_2^{(2)}}\dfrac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)1 \end{bmatrix}$$

$$= \left(\boldsymbol{f}'^{(2)} \odot \boldsymbol{\delta}^{(2)}\right)\widetilde{\boldsymbol{a}}^{(1)^T}$$

$$\boldsymbol{f}'^{(2)} \equiv \begin{bmatrix} \dfrac{\partial a_1^{(2)}}{\partial z_1^{(2)}} \\ \dfrac{\partial a_2^{(2)}}{\partial z_2^{(2)}} \end{bmatrix}$$

$$\boldsymbol{\delta}^{(2)} \equiv \begin{bmatrix} \dfrac{\partial C}{\partial a_1^{(2)}} \\ \dfrac{\partial C}{\partial a_2^{(2)}} \end{bmatrix}$$

$$= \begin{bmatrix} \displaystyle\sum_{k=1}^{4} \frac{\partial C}{\partial a_k^{(3)}}\frac{\partial a_k^{(3)}}{\partial z_k^{(3)}}W_{k1}^{(3)} \\ \displaystyle\sum_{k=1}^{4} \frac{\partial C}{\partial a_k^{(3)}}\frac{\partial a_k^{(3)}}{\partial z_k^{(3)}}W_{k2}^{(3)} \end{bmatrix}$$

$$= \boldsymbol{W}^{(3)^T}\left(\boldsymbol{f}'^{(3)} \odot \boldsymbol{\delta}^{(3)}\right)$$

Oaneed Nodkoasteen

Layer One

$$\frac{\partial C}{\partial \widetilde{\boldsymbol{W}}^{(1)}} \equiv \begin{bmatrix} \frac{\partial C}{\partial W_{11}^{(1)}} & \frac{\partial C}{\partial W_{12}^{(1)}} & \frac{\partial C}{\partial W_{13}^{(1)}} & \frac{\partial C}{\partial W_{14}^{(1)}} \\ \frac{\partial C}{\partial W_{21}^{(1)}} & \frac{\partial C}{\partial W_{22}^{(1)}} & \frac{\partial C}{\partial W_{23}^{(1)}} & \frac{\partial C}{\partial W_{24}^{(1)}} \\ \frac{\partial C}{\partial W_{31}^{(1)}} & \frac{\partial C}{\partial W_{32}^{(1)}} & \frac{\partial C}{\partial W_{33}^{(1)}} & \frac{\partial C}{\partial W_{34}^{(1)}} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{\partial C}{\partial a_1^{(1)}}\frac{\partial a_1^{(1)}}{\partial z_1^{(1)}}\right)a_1^{(0)} & \left(\frac{\partial C}{\partial a_1^{(1)}}\frac{\partial a_1^{(1)}}{\partial z_1^{(1)}}\right)a_2^{(0)} & \left(\frac{\partial C}{\partial a_1^{(1)}}\frac{\partial a_1^{(1)}}{\partial z_1^{(1)}}\right)a_3^{(0)} & \left(\frac{\partial C}{\partial a_1^{(1)}}\frac{\partial a_1^{(1)}}{\partial z_1^{(1)}}\right)1 \\ \left(\frac{\partial C}{\partial a_2^{(1)}}\frac{\partial a_2^{(1)}}{\partial z_2^{(1)}}\right)a_1^{(0)} & \left(\frac{\partial C}{\partial a_2^{(1)}}\frac{\partial a_2^{(1)}}{\partial z_2^{(1)}}\right)a_2^{(0)} & \left(\frac{\partial C}{\partial a_2^{(1)}}\frac{\partial a_2^{(1)}}{\partial z_2^{(1)}}\right)a_3^{(0)} & \left(\frac{\partial C}{\partial a_2^{(1)}}\frac{\partial a_2^{(1)}}{\partial z_2^{(1)}}\right)1 \\ \left(\frac{\partial C}{\partial a_3^{(1)}}\frac{\partial a_3^{(1)}}{\partial z_3^{(1)}}\right)a_1^{(0)} & \left(\frac{\partial C}{\partial a_3^{(1)}}\frac{\partial a_3^{(1)}}{\partial z_3^{(1)}}\right)a_2^{(0)} & \left(\frac{\partial C}{\partial a_3^{(1)}}\frac{\partial a_3^{(1)}}{\partial z_3^{(1)}}\right)a_3^{(0)} & \left(\frac{\partial C}{\partial a_3^{(1)}}\frac{\partial a_3^{(1)}}{\partial z_3^{(1)}}\right)1 \end{bmatrix}$$

$$= \left(\boldsymbol{f}'^{(1)} \odot \boldsymbol{\delta}^{(1)}\right)\widetilde{\boldsymbol{a}}^{(0)^T}$$

$$\boldsymbol{f}'^{(1)} \equiv \begin{bmatrix} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \\ \frac{\partial a_2^{(1)}}{\partial z_2^{(1)}} \\ \frac{\partial a_3^{(1)}}{\partial z_3^{(1)}} \end{bmatrix}$$

$$\boldsymbol{\delta}^{(1)} \equiv \begin{bmatrix} \frac{\partial C}{\partial a_1^{(1)}} \\ \frac{\partial C}{\partial a_2^{(1)}} \\ \frac{\partial C}{\partial a_3^{(1)}} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{2} \frac{\partial C}{\partial a_k^{(2)}}\frac{\partial a_k^{(2)}}{\partial z_k^{(2)}} W_{k1}^{(2)} \\ \sum_{k=1}^{2} \frac{\partial C}{\partial a_k^{(2)}}\frac{\partial a_k^{(2)}}{\partial z_k^{(2)}} W_{k2}^{(2)} \\ \sum_{k=1}^{2} \frac{\partial C}{\partial a_k^{(2)}}\frac{\partial a_k^{(2)}}{\partial z_k^{(2)}} W_{k3}^{(2)} \end{bmatrix}$$

$$= \boldsymbol{W}^{(2)^T}\left(\boldsymbol{f}'^{(2)} \odot \boldsymbol{\delta}^{(2)}\right)$$

Oarneed Noakoasteen

Table 8. Derivation of The He and Glorot Initialization Schemes

The Assumptions for Each Layer

The elements of the input/output are i.i.d (independent and identically distributed). (1-a)

The elements of the weight matrix are i.i.d (independent and identically distributed). Furthermore, it is assumed that the elements of the weight matrix are drawn from a zero-mean distribution. (1-b)

The elements of the input/output and of the weight matrix are independent of each other. (1-c)

The elements of the $\boldsymbol{\delta}$ are i.i.d (independent and identically distributed). (1-d)

The elements of the $\boldsymbol{\delta}$ and of the weight matrix are independent of each other. (1-e)

The probability distribution of the pre-non-linearity output, denoted by $p(z)$ is symmetric and centered at zero, i.e., it produces the probability of $\frac{1}{2}$ on both the positive and the negative parts of the z axis. (1-f)

The Variance of the Multiplication of Two Independent RVs

$$\begin{aligned} Var[xy] &= E[x^2y^2] - E[xy]^2 \\ &= E[x^2]E[y^2] - E[x]^2E[y]^2 \\ &= (Var[x] + E[x]^2)(Var[y] + E[y]^2) - E[x]^2E[y]^2 \\ &= Var[x]Var[y] + Var[x]E[y]^2 + Var[y]E[x]^2 \end{aligned} \quad \text{I-1}$$

The Law of the Unconscious Statistician

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)p(x)\,dx \quad \text{I-2}$$

The $\boldsymbol{a}^{(l)}$

$$\boldsymbol{a}^{(l)} = f\left(\widetilde{\boldsymbol{W}}^{(l)}\widetilde{\boldsymbol{a}}^{(l-1)}\right) \quad \text{(2-a)}$$

The k^{th} element of the $\boldsymbol{a}^{(l)}$

(2-a) $\rightarrow$

$$a_k^{(l)} = f\left(\sum_{i=1}^{(O_{l-1}+1)} w_{k,i}^{(l)} a_i^{(l-1)}\right) \quad \text{(2-b)}$$

Oameed Noakoasteen

The $\boldsymbol{\delta}^{(l)}$

$$\boldsymbol{\delta}^{(l)} = \boldsymbol{W}^{(l+1)^T}\left(\boldsymbol{f}'^{(l+1)} \odot \boldsymbol{\delta}^{(l+1)}\right) \tag{3-a}$$

The k^{th} element of the $\boldsymbol{\delta}^{(l)}$

(3-a) → $$\delta_k^{(l)} = \sum_{i=1}^{O_{l+1}} w_{i,k}^{(l+1)} \delta_i^{(l+1)} f'^{(l+1)}_i \tag{3-b}$$

Weight Distribution for the ReLU Non-linearity

The $f(z)$

$$f(z) \equiv \max(0, z) \tag{4-a}$$

(1-f), (I-2), (4-a) →

$$\begin{aligned} E[f^2] &= \int_{-\infty}^{+\infty} f^2(z) p(z)\, dz \\ &= \frac{1}{2} E[z^2] \end{aligned} \tag{4-b}$$

$$\begin{aligned} E\left[f'^2\right] &= \int_{-\infty}^{+\infty} f'^2(z) p(z)\, dz \\ &= \frac{1}{2} \end{aligned} \tag{4-c}$$

The Variance of $z_k^{(l)}$

(1-a) → Since the elements of $\boldsymbol{a}^{(l)}$ and, by extension, $\boldsymbol{z}^{(l)}$ are i.i.d, the k^{th} element of $\boldsymbol{a}^{(l)}$ and $\boldsymbol{z}^{(l)}$ are simply denoted by $a^{(l)}$ and $z^{(l)}$ instead of $a_k^{(l)}$ and $z_k^{(l)}$. (5-a)

(1-b), (1-c). (2-b), (5-a) →

$$\begin{aligned} E\left[z^{(l)}\right] &= E\left[\sum_{i=1}^{(O_{l-1}+1)} w_{k,i}^{(l)} a_i^{(l-1)}\right] \\ &= \sum_{i=1}^{(O_{l-1}+1)} E\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\ &= \sum_{i=1}^{(O_{l-1}+1)} E\left[w_{k,i}^{(l)}\right] E\left[a_i^{(l-1)}\right] \\ &= 0 \end{aligned} \tag{5-b}$$

Omeed Noakoasteen

$$
\begin{aligned}
Var\left[z^{(l)}\right] &= Var\left[\sum_{i=1}^{(O_{l-1}+1)} w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(O_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2 \left(a_i^{(l-1)}\right)^2\right] - E^2\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(O_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2\right] E\left[\left(a_i^{(l-1)}\right)^2\right] - E^2\left[w_{k,i}^{(l)}\right] E^2\left[a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(O_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2\right] E\left[\left(a_i^{(l-1)}\right)^2\right] \qquad (5\text{-c}) \\
&= \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] E\left[\left(a_i^{(l-1)}\right)^2\right] \\
&= \frac{1}{2} \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] E\left[\left(z_i^{(l-1)}\right)^2\right] \\
&= \frac{1}{2} \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] Var\left[z_i^{(l-1)}\right] \\
&= \frac{O_{l-1}}{2}\left(Var\left[w_k^{(l)}\right] Var\left[z^{(l-1)}\right]\right)
\end{aligned}
$$

From (5-c) it is evident that, to have $Var\left[z^{(l)}\right] = Var\left[z^{(l-1)}\right]$, the corresponding weight vector must satisfy $\frac{O_{l-1}}{2}\left(Var\left[w_k^{(l)}\right]\right) = 1$. (6)

The Variance of $\delta_k^{(l)}$

Since the elements of $\boldsymbol{\delta}^{(l)}$ are i.i.d, the k^{th} element of $\boldsymbol{\delta}^{(l)}$ is simply denoted by $\delta^{(l)}$ instead of $\delta_k^{(l)}$. (7-a)

$$
\begin{aligned}
E\left[\delta^{(l)}\right] &= E\left[\sum_{i=1}^{O_{l+1}} w_{i,k}^{(l+1)} \delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \\
&= \sum_{i=1}^{O_{l+1}} E\left[w_{i,k}^{(l+1)} \delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \qquad (7\text{-b}) \\
&= \sum_{i=1}^{O_{l+1}} E\left[w_{i,k}^{(l+1)}\right] E\left[\delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \\
&= 0
\end{aligned}
$$

(1-a), (1-b), (1-c), (I-1), (2-b), (4-b), (5-a), (5-b) →

(1-d) →

(1-e), (3-b), (7-a) →

Oameed Nakasteen

$$
\begin{aligned}
Var\left[\delta^{(l)}\right] &= Var\left[\sum_{i=1}^{O_{l+1}} w_{i,k}^{(l+1)} \delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \\
&= \sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)} \delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \\
&= \sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)}\right] Var\left[\delta_i^{(l+1)} {f'}_i^{(l+1)}\right] \\
&= \frac{1}{2}\sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)}\right] Var\left[\delta_i^{(l+1)}\right] \\
&= \frac{O_{l+1}}{2} Var\left[w_k^{(l+1)}\right] Var\left[\delta^{(l+1)}\right]
\end{aligned}
\qquad \text{(7-a), (7-b), (4-c), (3-b), (I-1), (1-c), (1-b), (1-a)} \rightarrow \text{(7-c)}
$$

From (7-c) it is evident that, to have $Var\left[\delta^{(l)}\right] = Var\left[\delta^{(l+1)}\right]$, the corresponding weight vector must satisfy $\frac{O_{l+1}}{2}\left(Var\left[w_k^{(l+1)}\right]\right) = 1$. This implies that for the l^{th} layer, $\frac{O_l}{2}\left(Var\left[w_k^{(l)}\right]\right) = 1$. (8)

The He Initialization

To keep constant the variance of $z_k^{(l)}$ as they propagate forward and the variance of $\delta_k^{(l)}$ as they propagate backward through the layers, either (6) or (8) can be used to initialize the weights of the l^{th} layer. (6), (8) → (9)

Normal Distribution

$$w^{(l)} \sim \mathcal{N}\left(0, \frac{2}{O_{l-1}}\right)$$

(9) →

Uniform Distribution

$$w^{(l)} \sim U\left(-\sqrt{\frac{6}{O_{l-1}}}, \sqrt{\frac{6}{O_{l-1}}}\right)$$

Weight Distribution for the Hyperbolic Tangent Non-linearity

The $f(z)$

$$f(z) \equiv \tanh(z) \tag{10-a}$$

Tylor Expansion

$$\tanh(z) \approx z - \cdots \tag{10-b}$$

Oameed Noakoasteen

$$
\begin{aligned}
E[f^2] &= \int_{-\infty}^{+\infty} f^2(z)p(z)\,dz \\
&= \int_{-\infty}^{+\infty} z^2 p(z)\,dz \\
&= E[z^2]
\end{aligned}
\tag{10-c}
$$

(1-f), (I-2), (10-a), (10-b) →

$$
\begin{aligned}
E\left[f'^2\right] &= \int_{-\infty}^{+\infty} \left(f'^2(z)\right)p(z)\,dz \\
&= \int_{-\infty}^{+\infty} p(z)\,dz \\
&= 1
\end{aligned}
\tag{10-d}
$$

The Variance of $z_k^{(l)}$

(1-a) → Since the elements of $\boldsymbol{a}^{(l)}$ and, by extension, $\boldsymbol{z}^{(l)}$ are i.i.d, the k^{th} element of $\boldsymbol{a}^{(l)}$ and $\boldsymbol{z}^{(l)}$ are simply denoted by $a^{(l)}$ and $z^{(l)}$ instead of $a_k^{(l)}$ and $z_k^{(l)}$. (11-a)

$$
\begin{aligned}
E\left[z^{(l)}\right] &= E\left[\sum_{i=1}^{(o_{l-1}+1)} w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} E\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} E\left[w_{k,i}^{(l)}\right] E\left[a_i^{(l-1)}\right] \\
&= 0
\end{aligned}
\tag{11-b}
$$

(1-b), (1-c). (2-b), (11-a) →

$$
\begin{aligned}
Var\left[z^{(l)}\right] &= Var\left[\sum_{i=1}^{(o_{l-1}+1)} w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} Var\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2 \left(a_i^{(l-1)}\right)^2\right] - E^2\left[w_{k,i}^{(l)} a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2\right] E\left[\left(a_i^{(l-1)}\right)^2\right] - E^2\left[w_{k,i}^{(l)}\right] E^2\left[a_i^{(l-1)}\right] \\
&= \sum_{i=1}^{(o_{l-1}+1)} E\left[\left(w_{k,i}^{(l)}\right)^2\right] E\left[\left(a_i^{(l-1)}\right)^2\right]
\end{aligned}
\tag{11-c}
$$

(1-a), (1-b), (1-c), (I-1), (2-b), (10-c), (11-a), (11-b) →

Oameed Noakoasteen

$$= \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] E\left[\left(a_i^{(l-1)}\right)^2\right]$$

$$= \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] E\left[\left(z_i^{(l-1)}\right)^2\right]$$

$$= \sum_{i=1}^{(O_{l-1}+1)} Var\left[w_{k,i}^{(l)}\right] Var\left[z_i^{(l-1)}\right]$$

$$= O_{l-1}\left(Var\left[w_k^{(l)}\right] Var\left[z^{(l-1)}\right]\right)$$

From (11-c) it is evident that, to have $Var\left[z^{(l)}\right] = Var\left[z^{(l-1)}\right]$, the corresponding weight vector must satisfy $O_{l-1}\left(Var\left[w_k^{(l)}\right]\right) = 1$. (12)

The Variance of $\delta_k^{(l)}$

(1-d) → Since the elements of $\boldsymbol{\delta}^{(l)}$ are i.i.d, the k^{th} element of $\boldsymbol{\delta}^{(l)}$ is simply denoted by $\delta^{(l)}$ instead of $\delta_k^{(l)}$. (13-a)

$$E\left[\delta^{(l)}\right] = E\left[\sum_{i=1}^{O_{l+1}} w_{i,k}^{(l+1)} \delta_i^{(l+1)} f'_i{}^{(l+1)}\right]$$

(1-e), (3-b), (13-a) →

$$= \sum_{i=1}^{O_{l+1}} E\left[w_{i,k}^{(l+1)} \delta_i^{(l+1)} f'_i{}^{(l+1)}\right] \quad (13\text{-b})$$

$$= \sum_{i=1}^{O_{l+1}} E\left[w_{i,k}^{(l+1)}\right] E\left[\delta_i^{(l+1)} f'_i{}^{(l+1)}\right]$$

$$= 0$$

$$Var\left[\delta^{(l)}\right] = Var\left[\sum_{i=1}^{O_{l+1}} w_{i,k}^{(l+1)} \delta_i^{(l+1)} f'_i{}^{(l+1)}\right]$$

(1-a), (1-b), (1-c), (1-1), (3-b), (10-d), (13-b) →

$$= \sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)} \delta_i^{(l+1)} f'_i{}^{(l+1)}\right]$$

$$= \sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)}\right] Var\left[\delta_i^{(l+1)} f'_i{}^{(l+1)}\right] \quad (13\text{-c})$$

$$= \sum_{i=1}^{O_{l+1}} Var\left[w_{i,k}^{(l+1)}\right] Var\left[\delta_i^{(l+1)}\right]$$

$$= O_{l+1} Var\left[w_k^{(l+1)}\right] Var\left[\delta^{(l+1)}\right]$$

Oameed Noakoasteen

From (13-c) it is evident that, to have $Var[\delta^{(l)}] = Var[\delta^{(l+1)}]$, the corresponding weight vector must satisfy $O_{l+1}\left(Var\left[w_k^{(l+1)}\right]\right) = 1$. (14)
This implies that for the l^{th} layer, $O_l\left(Var\left[w_k^{(l)}\right]\right) = 1$.

The Glorot Initialization

To keep constant the variance of $a_k^{(l)}$ as they propagate forward and the variance of $\delta_k^{(l)}$ as they propagate backward through the layers, $Var\left[w_k^{(l)}\right] = \frac{2}{O_l + O_{l-1}}$ can be used to initialize the weights of the l^{th} layer. This satisfies both (12) and (14) when the l^{th} layer is connected to layers of the same size at both its input and its output. (15)

(12), (14) →

Normal Distribution

$$w^{(l)} \sim \mathcal{N}\left(0, \frac{2}{O_l + O_{l-1}}\right)$$

(15) →

Uniform Distribution

$$w^{(l)} \sim U\left(-\sqrt{\frac{6}{O_l + O_{l-1}}}, \sqrt{\frac{6}{O_l + O_{l-1}}}\right)$$

Oameed Noakoasteen

Part III: Appendices

Oameed Noakoasteen

Oameed Noakoasteen

Appendix I - A Medley of Concepts

Table 9. Derivation of the Bias Variance Decomposition for MSE

Some Concepts

In the following derivation, the expectation is computed with respect to all examples $(\boldsymbol{x}, y)$ that the model f_D encounters in the inference phase.

The expected value of the prediction for a test datapoint $\boldsymbol{x}$ is denoted by $\overline{f_D}(\boldsymbol{x})$.

The expected value of the target for a test datapoint $\boldsymbol{x}$ is denoted by $\bar{y}(\boldsymbol{x})$.

$$\begin{aligned} E\left[\left(y - f_D(\boldsymbol{x})\right)^2\right] &= E\left[\left(\left(y - \overline{f_D}(\boldsymbol{x})\right) - \left(f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)\right)^2\right] \\ &= E\left[\left(y - \overline{f_D}(\boldsymbol{x})\right)^2\right] + E\left[\left(f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] \\ &- 2E\left[\left(y - \overline{f_D}(\boldsymbol{x})\right)\left(f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)\right] \end{aligned} \tag{1}$$

The two expressions $y - \overline{f_D}(\boldsymbol{x})$ and $f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})$ are statistically independent. Furthermore, $E\left[f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right] = 0$. (2)

(1), (2) $\rightarrow$ $$E\left[\left(y - f_D(\boldsymbol{x})\right)^2\right] = E\left[\left(y - \overline{f_D}(\boldsymbol{x})\right)^2\right] + E\left[\left(f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] \tag{3}$$

$$\begin{aligned} E\left[\left(y - \overline{f_D}(\boldsymbol{x})\right)^2\right] &= E\left[\left(\left(y - \bar{y}(\boldsymbol{x})\right) + \left(\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)\right)^2\right] \\ &= E\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + E\left[\left(\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] \\ &+ 2E\left[\left(y - \bar{y}(\boldsymbol{x})\right)\left(\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)\right] \end{aligned} \tag{4}$$

The two expressions $y - \bar{y}(\boldsymbol{x})$ and $\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})$ are statistically independent. Furthermore, $E[y - \bar{y}(\boldsymbol{x})] = 0$. (5)

(4), (5) $\rightarrow$ $$E\left[\left(y - \overline{f_D}(\boldsymbol{x})\right)^2\right] = E\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + E\left[\left(\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] \tag{6}$$

(3), (6) $\rightarrow$ $$E\left[\left(y - f_D(\boldsymbol{x})\right)^2\right] = E\left[\left(\bar{y}(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] + E\left[\left(f_D(\boldsymbol{x}) - \overline{f_D}(\boldsymbol{x})\right)^2\right] + E\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right]$$

Omeed Nookasteen

Table 10. Derivation of the mean and variance of predictions in Bayesian inference

The Posterior Distribution of $\widetilde{\boldsymbol{w}}$

$$p\left(\widetilde{\boldsymbol{w}}|\widetilde{\boldsymbol{X}},\boldsymbol{y}\right)=\mathcal{N}\left(\boldsymbol{\mu}=\boldsymbol{A}\boldsymbol{b},\boldsymbol{\Sigma}=\sigma^2\boldsymbol{A}\right) \tag{1}$$

$$\boldsymbol{A}\equiv\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}+\sigma^2\left(\boldsymbol{\Sigma}^{prior}\right)^{-1}\right)^{-1}$$
$$\boldsymbol{b}\equiv\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$$

Expected Value and Variance of a Single-Datapoint Prediction

$$\hat{y}=\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}} \tag{2}$$

(2) $\rightarrow$

$$E[\hat{y}]=E^T[\widetilde{\boldsymbol{w}}]\widetilde{\boldsymbol{x}} \tag{3}$$

(2), (3) $\rightarrow$

$$\begin{aligned} Var[\hat{y}] &\equiv E[(\hat{y}-E[\hat{y}])(\hat{y}-E[\hat{y}])] \\ &= E[\hat{y}^2]-E^2[\hat{y}] \\ &= E[\hat{y}^T\hat{y}]-E^T[\hat{y}]E[\hat{y}] \\ &= E[\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{w}}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}]-\widetilde{\boldsymbol{x}}^TE[\widetilde{\boldsymbol{w}}]E^T[\widetilde{\boldsymbol{w}}]\widetilde{\boldsymbol{x}} \\ &= \widetilde{\boldsymbol{x}}^TE[\widetilde{\boldsymbol{w}}\widetilde{\boldsymbol{w}}^T]\widetilde{\boldsymbol{x}}-\widetilde{\boldsymbol{x}}^TE[\widetilde{\boldsymbol{w}}]E^T[\widetilde{\boldsymbol{w}}]\widetilde{\boldsymbol{x}} \\ &= \widetilde{\boldsymbol{x}}^T(E[\widetilde{\boldsymbol{w}}\widetilde{\boldsymbol{w}}^T]-E[\widetilde{\boldsymbol{w}}]E^T[\widetilde{\boldsymbol{w}}])\widetilde{\boldsymbol{x}} \\ &= \widetilde{\boldsymbol{x}}^TK_{\widetilde{\boldsymbol{w}}\widetilde{\boldsymbol{w}}}\widetilde{\boldsymbol{x}} \end{aligned} \tag{4}$$

(1), (3), (4) $\rightarrow$

$$E[\hat{y}]=\boldsymbol{b}^T\boldsymbol{A}\widetilde{\boldsymbol{x}}$$
$$Var[\hat{y}]=\sigma^2\widetilde{\boldsymbol{x}}^T\boldsymbol{A}\widetilde{\boldsymbol{x}}$$

Oameed Noakoasteen

Appendix II - Useful Math

Probability & Statistics

- Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Bayes' Theorem

$$P(A|B) = \left[\frac{1}{P(B)} P(B|A)\right] P(A)$$

In the above expression, $P(A)$ is the **Prior** probability of observing A without any given conditions; $P(B|A)$ is the **Likelihood**; $\frac{1}{P(B)}$ is the **Evidence**; And, $P(A|B)$ is the **Posterior** probability of A given B.

- Probability Distributions

Multivariate Gaussian

$$\boldsymbol{x} \in \mathbb{R}^n \sim \mathcal{N}\left(\boldsymbol{\mu}_{n\times 1}, \boldsymbol{\Sigma}_{n\times n}\right)$$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Bernoulli

$$x \in \{1, 0\} \sim \text{Bern}(\mathfrak{p})$$

$$p(x) = \mathfrak{p}^x (1-\mathfrak{p})^{1-x}$$

- Covariance & Correlation

Random Variables x and y

$$\text{Cov}(x, y) \equiv \mathbb{E}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right]$$

Variance

$$\sigma_x^2 \equiv \text{Cov}(x, x)$$

Random Vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) \equiv \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T\right]$$

Covariance Matrix

$$K_{\boldsymbol{xx}} \equiv \text{Cov}(\boldsymbol{x}, \boldsymbol{x})$$

Pearson Correlation Coefficients

$$\rho_{\boldsymbol{xy}} \equiv \mathbb{E}\left[\left(\frac{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]}{\sigma_{\boldsymbol{x}}}\right)\left(\frac{\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}]}{\sigma_{\boldsymbol{y}}}\right)^T\right]$$

Auto-Correlation Matrix

$$R_{\boldsymbol{xx}} \equiv \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right]$$

- $\frac{1}{m} \boldsymbol{X}^T_{m\times n} \boldsymbol{X}_{m\times n}$

Depending on how $\boldsymbol{X}$ (the design matrix) is pre-processed, $\frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}$ can be the Auto-Correlation Matrix (no pre-processing) or the Covariance Matrix (subtracting the mean of each column from that column) or Pearson Correlation Coefficients Matrix (subtracting the mean and dividing by standard deviation). The $\boldsymbol{X}\boldsymbol{X}^T$ is referred to as the Gram Matrix.

Information Theory

- Entropy

$$\begin{aligned} H(p) &\equiv -\sum_{x\in X} p(x) \log p(x) \\ &= \mathbb{E}\left[-\log p(X)\right] \end{aligned}$$

- Cross-Entropy

$$\begin{aligned} H(p, q) &\equiv -\sum_{x\in X} p(x) \log q(x) \\ &= \mathbb{E}_p\left[-\log q(X)\right] \end{aligned}$$

- KL Divergence

$$\begin{aligned} D_{KL}(p||q) &\equiv \sum_{x\in X} p(x) \log \frac{p(x)}{q(x)} \\ &= \mathbb{E}\left[\log \frac{p(X)}{q(X)}\right] \end{aligned}$$

Special Functions

- Square Exponential (Gaussian) Kernel

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2}$$

- Indicator Function

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Oameed Noakoasteen

■ Singular Value Decomposition

- $\boldsymbol{A}_{M\times N} = \boldsymbol{U}_{M\times M}\boldsymbol{\Sigma}_{M\times N}\boldsymbol{V}^T_{N\times N}$

  Left Singular Vectors

  $$\boldsymbol{U}_{M\times M} \equiv \left[\boldsymbol{u}^{(1)}_{M\times 1}\cdots\boldsymbol{u}^{(M)}_{M\times 1}\right]$$
  $$\begin{aligned}\boldsymbol{U}\boldsymbol{U}^T &= \boldsymbol{U}^T\boldsymbol{U}\\ &= \boldsymbol{I}_{M\times M}\end{aligned}$$

  Singular Values

  Diagonal elements of $\boldsymbol{\Sigma}_{M\times N}$ are $\{\sigma_1,\cdots,\sigma_r\}$ in which $r \equiv \min(M,N)$. Non-diagonal elements of $\boldsymbol{\Sigma}_{M\times N}$ are zero.

  Right Singular Vectors

  $$\boldsymbol{V}_{N\times N} \equiv \left[\boldsymbol{v}^{(1)}_{N\times 1}\cdots\boldsymbol{v}^{(N)}_{N\times 1}\right]$$
  $$\begin{aligned}\boldsymbol{V}\boldsymbol{V}^T &= \boldsymbol{V}^T\boldsymbol{V}\\ &= \boldsymbol{I}_{N\times N}\end{aligned}$$

- Diagonalization of A

  $$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma},\ {\boldsymbol{u}^{(i)}}^T\boldsymbol{A}\boldsymbol{v}^{(i)} = \sigma_i$$

- Low-Rank Approximation of A

  $$\boldsymbol{A} = \sum_{i=1}^{k}\sigma_i\boldsymbol{u}^{(i)}{\boldsymbol{v}^{(i)}}^T$$

- Pseudo-Inverse of A

  $$\boldsymbol{A}^{-1}_{N\times M} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$$

- Compact (Economy) SVD

  $$\boldsymbol{A}_{M\times N} = \boldsymbol{U}_{M\times r}\boldsymbol{\Sigma}_{r\times r}\boldsymbol{V}^T_{N\times r}$$

- Eigenvalue Decomposition

  $$\boldsymbol{A}_{N\times N} = \boldsymbol{Q}_{N\times N}\boldsymbol{\Lambda}_{N\times N}\boldsymbol{Q}^T_{N\times N}$$

Oameed Noakoasteen